AFRL-OSR-VA-TR-2015-0067

TOWARDS A GENERAL THEORY OF COUNTERDECEPTION

Scott Craver
RESEARCH FOUNDATION OF STATE UNIVERSITY OF NEW YORK THE

02/20/2015
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTC
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
02/06/2015	FINAL	09/14/2009 to 09/14/2014

4. TITLE AND SUBTITLE

TOWARD A GENERAL THEORY OF COUNTERDECEPTION

5a. CONTRACT NUMBER

5b. GRANT NUMBER

FA9550-09-1-0666

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

SCOTT CRAVER

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

BINGHAMTON UNIVERSITY
4400 Vestal Pkwy E
Binghamton NY 13902-6000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
801 N Randolph St., Rm. 732
Arlington VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION A -- APPROVED FOR PUBLIC RELEASE

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Research in the fields of information hiding and digital forensics has introduced a new kind of problem, the deception problem, which is beyond the theoretical scope of mainstream cryptography. In this problem, one party attempts to detect malicious behavior, while the other party seeks to evade or fool a detection algorithm. This is at its core a classification or signal processing problem, except we are impeded not by noise, but by an intelligent adversary. These problems include many existing problems from virus detection to the detection of network attacks. The effort of an adversary to find malicious input that evades detection has yet to be quantified in the same way as brute-forcing an encryption primitive.
This effort explored an adversarial version of detection and estimation theory, to uncover fundamental theoretical limits on an adversary's performance, or a detector's power in uncovering calculated deception, as well as limiting information leakage from its own outputs. Ultimately this effort seeks to answer the theoretical question of which party in a deception problem has the upper hand—the detector or the deceiver—subject to a set of initial conditions determined by the specific problem.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(Include area code)*

FINAL PROJECT REPORT

Towards a General Theory of Counterdeception

Award No. FA9550-09-1-0666

Awarded to THE RESEARCH FOUNDATION OFSTATE UNIVERSITY OF NEW YORK (CAGE code 3GRK1)

SUNY AT BINGHAMTON

4400 Vestal Pkwy E

Binghamton NY 13902-6000

Principal Investigator: Dr. Scott A. Craver

Assistant Professor, Department of Electrical and Computer Engineering

University of Binghamton

Binghamton, NY 13902-6000

Abstract

Research in the fields of information hiding and digital forensics has introduced a new kind of problem, the deception problem, which is beyond the theoretical scope of mainstream cryptography. In this problem, one party attempts to detect malicious behavior, while the other party seeks to evade or fool a detection algorithm. This is at its core a classification or signal processing problem, except we are impeded not by noise, but by an intelligent adversary. These problems include many existing problems from virus detection to the detection of network attacks. The effort of an adversary to find malicious input that evades detection has yet to be quantified in the same way as brute-forcing an encryption primitive.

This effort explored an adversarial version of detection and estimation theory, to uncover fundamental theoretical limits on an adversary's performance, or a detector's power in uncovering calculated deception, as well as limiting information leakage from its own outputs. Ultimately this effort seeks to answer the theoretical question of which party in a deception problem has the upper hand—the detector or the deceiver—subject to a set of initial conditions determined by the specific problem.

Overview

Cryptography has effectively solved the problem of secure communication with guarantees of confidentiality and integrity. Unfortunately, today's security problems are not so simple, and do not neatly fit the model of two parties securing a channel.

Research in the fields of information hiding and digital forensics has introduced a new kind of problem: the deception problem. In this problem, one party attempts to detect malicious behavior, while the other party seeks to evade or fool a detection algorithm. This is at its core a classification or signal processing problem, except we are impeded not by noise, but by an intelligent adversary. Detecting a hidden threat is an act of *counterdeception;* purposefully evading this detector is a *deception* tactic.

Deception problems include many vexing security problems which have no clear solution for either the detector or the deceiver:

- Virus detection and intrusion detection
- Tamper-evidence of images
- Anonymity and pseudonymity protocols
- Steganography, or data hiding
- Detection of covert channels
- Face recognition and voice recognition
- Watermarking of multimedia

Unlike the classical problem of secure communication, deception problems have not converged to a "win" for one party or the other. They are cat-and-mouse games, in which detectors are improved and deceivers adapt to new detectors. It is not yet clear who ultimately has the upper hand in any of these problems. This is a sharp contrast to cryptology, in which perfect secrecy is not only provable, but realistically achievable. *Deception problems do not yet benefit from this level of theoretical development.*

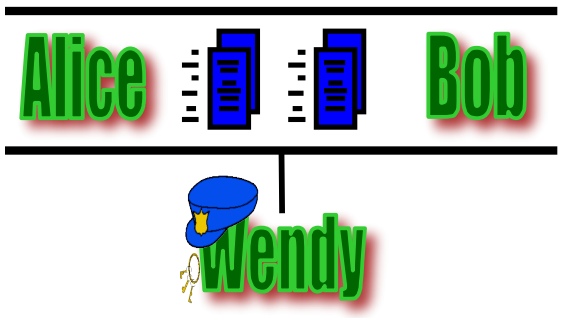

Figure 1: The prisoners' problem, a standard model for covert communication. Alice and Bob must conceal messages within innocent cover-data, while a warden analyzes all traffic.

Even the basic ability to characterize a detector's brute-forcing time is not possible with our current lack of formalism. In mainstream cryptographic problems, a communications line is encrypted with a secret key that, absent any flaws in the cipher or its implementation, must be guessed by brute force search. The complexity of brute force search is simple to compute, because keys are drawn with a fixed distribution, often uniform, and from a well-defined key space; furthermore an incorrectly guessed key does not provide any partial information to an attacker, beyond reducing the remaining search space by one guess. The complexity is often expressed using the *guessing entropy*, $H_G(X)$

$= 1+log_2(\sum_k k \cdot Pr[x_k])$, where $Pr[x_k] \geq Pr[x_{k+1}]$. This is simply the expected amount of effort to break a system, expressed on a logarithmic scale for consistency with logical key lengths: if a 128-bit key is chosen uniformly, its guessing entropy is 128.

In adversarial detection problems, the guessing entropy no longer characterizes the strength or vulnerability of the security system. Instead of guessing one key, an adversary must guess any desirable input that reverses the detector. Secondly, unlike an encryption algorithm, a detector's response to previous inputs leaks information about the performance of future inputs. The detector's algorithm may be characterized by a small set of guesses. There is presently no "reversing entropy" for a detection algorithm, or well-defined limit on its resistance to attack.

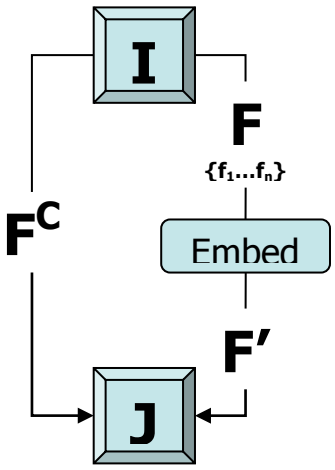

Figure 2: a block diagram of a typical feature-based information hiding algorithm.

Many counterdeception practices follow a standard formula of extracting statistical features from suspect data, which are fed into a simple detection algorithm. Data hiding techniques such as multimedia watermarking and fingerprinting will extract image features and modify them to encode data; in detection, those features are

extracted from a suspect and tested for similarity. A similar extraction of statistical features is used to detect covert data hiding, and algorithms to embed secret messages in multimedia data also seek to preserve the data's natural statistics. This "feature space" model can be analyzed in the abstract, and we can predict conditions that lead to security problems. For example, any statistical dependence between the statistical features F and remaining information F^C makes covert communication detectable. Even if a detector is not explicitly designed according to this framework of feature subspace extraction, it may be implicitly modelable in this way, and subject to attack according to the failures of that model.

The goal of this project is developing a mathematical foundation for "counterdeception" problems. In a counterdeception problem, one party seeks to detect unauthorized behavior while another seeks to evade detection. Many difficult problems in information security fit this template, including intrusion and virus detection, steganography and watermarking, forensics and biometrics. In particular, a great deal of security research at Binghamton University falls under the umbrella of counterdeception. Despite its importance in security, these problems lack a theoretical foundation such as that enjoyed by mainstream cryptography. For example, there is no easy way to quantify the time it takes an

adversary to brute-force a detector by probing it with experimental inputs, or even to build a watermark detector that takes an exponential amount of time to break, as a function of its key length. The state of detection is similar to the state of cryptography in the early 20th century: adversaries invent new technologies and attacks, which are quickly defeated by new adaptations by the other side. This cat-and-mouse game converged for cryptographic problems by the 1970s, due to several decades of mathematical and scientific progress; the hard problems of adversarial detection are likewise in need of a mathematical foundation.

Technical status of reporting period (Sept 2009-Sept 2014)

Our project has focused on several problems in reverse-engineering generic detection regions. First, we have focused on the problem of characterizing the "guessing entropy" of a detection region, in other words the approximate effort needed by an adversary to determine its shape by acquiring points on its surface. Classical detection regions are often very easy to reverse engineer: a secret watermark of n samples detected by correlation will have a detection region whose boundary is an n-dimensional hyperplane, and n points on this surface are sufficient to reverse engineer the entire detector. Contrast this with cryptography, in which an n-bit key requires an amount of brute-forcing effort exponential in n.

This ability to rapidly defeat a detector is a systemic problem in watermarking; there have been attempts to impede brute forcing by ad-hoc methods such as randomizing the detection boundary or choosing an algorithm whose detection region has an unusual shape, but robustness requirements of a watermark detector force the detection region to be fairly contiguous and well-behaved. Our first challenge was the creation of a detection region requiring an exponential amount of effort to defeat.

We have experimented with a technique called "random dot watermarking," which creates a detector whose output is the logical OR of many random detectors with very high thresholds. This produces a detection region made of many small random discs in the signal sphere. We have shown that this produces a region with exponential cracking effort, while guaranteeing the usual requirements for a watermark detector: that it be robust and have low false-alarm and miss rates. We did not, however, address the computational problem of quickly determining if an image is watermarked, or how to watermark quickly. Our current challenge is turning this result into a practical detector, and moreover applying the same result outside of digital watermarking.

A second focus of our effort is derivation of an optimal boundary rolloff for a secure detection region. It has long been known that

we can randomize a detection boundary to impede reverse-engineering, outputting a random result when an image is too close to the threshold of detection. This has not been shown to substantially increase brute-forcing time, however, and in fact it can only slow it by a constant factor: without randomization an adversary can find a boundary point by binary search; with randomization an adversary can still estimate a boundary point by averaging many samples.

A third issue we have explored is the theoretical detectability of steganographic embedding. It is known that if embedding alters the distribution of a cover text, the difference in distribution is ultimately detectable with sufficient samples; to prevent this, one must embed at a rate that asymptotically drops to zero. However, we can prevent this problem with subset selection. In lay terms, we grab a subset of data that already looks like tampered data, and then tamper with it, preserving the overall distribution of the signal. Given an embedding method, we can derive an artificial probability distribution **Q** that is unchanged by embedding. Then, given natural data with distribution **P**, we can find a constant a such that **aQ<P**. This is the maximal size of a chosen subset that will appear to have distribution **Q**. Technically, choosing that subset allows us to embed with an asymptotically constant rate.

A fourth topic that has been explored within the final two years of the reporting period is extension of our approach from information hiding to attacking biometric detectors. Specific findings are discussed below, in the technical status of our final reporting period. In short, we have discovered generic design flaws in common detection approaches used for biometric detectors, which allow exploitation of both speech recognition and face recognition systems.

One further accomplishment made under this grant is the development of new forensic methods for detecting zoom lenses in digital SLR (single lens reflex) cameras. While we primarily focus on the detection problem of watermarking and information hiding, forensic detection is within our umbrella of counterdeception problems. Working entirely within the watermarking domain is a liability because it has unique constraints: a watermark detector attempts to detect a secret signal of our own manufacture, using a detection region that can be any shape, as long as it satisfies the basic requirements of false-alarm and miss rates, and watermark robustness. In contrast, a forensic detector must detect a signal that is present in data, and the optimal detection region is beyond our control. Securing a forensic detector is a much harder problem, owing to the lack of freedom in design.

Technical status of final reporting period (Sept 2013-Sept 2014)

Several attacks have been discovered on biometric detectors since the previous report, specifically face and voice recognition systems. These include general attacks that exploit the multi-modal nature of implicit statistical models (for example, modeling a set of voice features as a Gaussian mixture,) which allow the injection or overwriting of mixtures with data that can trigger a misclassification in carefully engineered circumstances.

We have also identified a specific bug in common face recognition software and existing software libraries that allow a negative number attack on face recognition systems. These exploit the tendency of certain formulas used in detection, such as Chi-squared comparison of probability distributions, to assume nonnegative values; the assumption of nonnegativity is rarely enforced, and a negative value injected into a database can cause unusual effects. We demonstrated an attack in which a few bits are altered in a face recognition database---a floating point value 0 is replaced with a floating point value -0.0000001, for example---which causes an attacker to be classified as a target user whenever a certain printed gradient pattern is displayed to a face recognition system. These vulnerabilities are significant because they cause no

unusual behavior except when triggered by a specific signal from an attacker.

A third development in the final year of this project is the development of a steganographic virtual operating system. This conceals a user's data in various files on a computer system, like existing stegnographic file systems, but with a less conspicuous footprint, and augmenting said filesystem with a user interface and applications, creating an entire environment for viewing and manipulating hidden data.

Personnel Supported

Scott A. Craver, principal investigator.
Associate Professor, Department of Electrical and Computer Engineering
University of Binghamton

Idris Atakli, graduate student
Jun Yu, graduate student.
Enping Li, graduate student.
Caleb Serafy, graduate student
Alireza Farrokh baroughi, graduate student
Elan Ashendorf, graduate student

Mohammed Faizan Mohsin, undergraduate student
Virginia Li, undergraduate student

Publications:

1. A. Farrokh baroughi and S.A. Craver, “The Krusty the Clown attack on model-based speaker recognition systems.” In IS&T/SPIE Electronic Imaging (2015)

2. E. Ashendorf and S.A. Craver, “Design of a steganographic virtual operating system.” In IS&T/SPIE Electronic Imaging (2015)

3. A. Farrokh baroughi and S.A. Craver, “Additive Attacks on Speaker Recognition.” In IS&T/SPIE Electronic Imaging (2014)

4. J. Yu and S.A. Craver, “A fast, automatic camera image stabilization benchmarking scheme.” In IS&T/SPIE Electronic Imaging (2012)

5. E. Li and S.A. Craver, “A Square Root Law for Active Wardens.” Proceedings of the thirteenth ACM multimedia workshop on Multimedia and security (2011)

6. J. Yu and S.A. Craver, “Toward the identification of DSLR lenses by chromatic aberration.” in Media Forensics and

Security 2011, Proceedings of the SPIE (2011).

7. J. Yu and S.A. Craver, “Subset Selection Circumvents the Square Root Law.” in Media Forensics and Security 2010, Proceedings of the SPIE, Volume 7541 (2010).

8. J. Yu and S.A. Craver, “Reverse-engineering a watermark detector based on a more precise model.” IS&T/SPIE Electronic Imaging (2010).

There are two papers currently in preparation based on results collected in the final year of the project.

New inventions or patent disclosures:

There have been no invention or patent disclosures.